PB83-191478

Validation of Data Obtained during Exposure of Human Volunteers to Air Pollutants

Rockwell International
Chapel Hill, NC

Prepared for

Health Effects Research Lab.
Research Triangle Park, NC

1982

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/J-82-179	2. JOURNAL ARTICLE	3. R PB83-191478
4. TITLE AND SUBTITLE Validation of Data Obtained During Exposure of Human Volunteers to Air Pollutants		5. REPORT DATE 1982 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Donald E. Glover, Robert Hall, Richard Trilling, and Arthur W. Coston		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Rockwell International, Environmental Monitoring Services 800 Eastowne Drive, Suite 200 Chapel Hill, North Carolina 27514		10. PROGRAM ELEMENT NO. A9RA1B 11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Office of Research and Development Health Effects Research Laboratory US Environmental Protection Agency Research Triangle Park, NC 27711		13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA-600/11

15. SUPPLEMENTARY NOTES
Published In: Computers and Biomedical Research 15:240-249 (1982)

16. ABSTRACT

The Environmental Protection Agency (EPA) conducts research at their Clinical Environmental Laboratory to determine the effects of gaseous and particulate pollutants on volunteer human subjects. Procedures for the validation and management of data acquired from this facility are described, which ensure that data are both scientifically sound and legally defensible. The procedures conform to all applicable EPA quality control guidelines and provide a "chain-of-custody" record for the data capable of demonstrating that no unauthorized changes or deletions have occurred.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

18. DISTRIBUTION STATEMENT Release to Public	19. SECURITY CLASS (This Report) Unclassified	21. NO. OF PAGES 11
	20. SECURITY CLASS (This page) Unclassified	22. PRICE

EPA Form 2770-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

COMPUTERS AND BIOMEDICAL RESEARCH 15, 240–249 (1982)

EPA-600/J-82-179
JOURNAL ARTICLE

Validation of Data Obtained During Exposure of Human Volunteers to Air Pollutants

DONALD E. GLOVER, ROBERT. G. HALL, RICHARD J. TRILLING, AND ARTHUR W. COSTON

Rockwell International, Environmental Monitoring & Services Center, 800 Eastowne Drive, Suite 200, Chapel Hill, North Carolina 27514

AND

RALPH W. STACY

Clinical Studies Branch, Air Toxicology Division, Health Effects Research Laboratory, U. S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711

Received October 5, 1981

The Environmental Protection Agency (EPA) conducts research at their Clinical Environmental Laboratory to determine the effects of gaseous and particulate pollutants on volunteer human subjects. Procedures for the validation and management of data acquired from this facility are described, which ensure that data are both scientifically sound and legally defensible. The procedures conform to all applicable EPA quality control guidelines and provide a "chain-of-custody" record for the data, capable of demonstrating that no unauthorized changes or deletions have occurred.

INTRODUCTION

To provide human exposure data for use in setting air quality standards, the Environmental Protection Agency has designed and built a Clinical Environmental Laboratory (CEL) on the campus of the University of North Carolina at Chapel Hill (*1*). This facility consists of two 84 m^3 (3000 ft^3) environmental laboratories which are used to expose volunteer human subjects to constant or varying concentrations of SO_2, CO, NO_2, O_3, and water soluble aerosols. Changes in pulmonary and cardiovascular functions in response to the pollutant exposures are recorded using a computer-assisted Physiological Data Acquisition System (PDAS). A computer-based Aerosol Control System (*2*) and Gas Control System (*3*) regulate the test environment.

To ensure that data acquired from the CEL are legally defensible and scientifically acceptable, EPA requires that this data be subjected to "validation," defined by EPA as "The process whereby data are filtered and accepted or rejected based on a set of criteria" (*4*). Since the volume of data

0010-4809/82/030240-10$02.00/0
Copyright © 1982 by Academic Press, Inc.
All rights of reproduction in any form reserved.

U.S. Government is authorized to reproduce and sell this report. Permission for further reproduction must be obtained from the copyright owner.

from the CEL facility (10^6–10^7 data items per year) makes complete manual validation impractical, a computerized system was developed which "flags" (according to user-defined criteria) the smaller subset of suspicious data requiring careful manual inspection by the Principal Investigator (PI) of the study. Because the PI has the ultimate responsibility for the integrity of the data, it was necessary that this validation system not limit his/her ability to delete, flag, or modify data according to his/her scientific and professional judgment. To ensure that the data are legally defensible, however, the validation system was designed to maintain an "audit trail" of all changes made to the data base by the PI. This audit trail identifies the date and time the change was made, the "author" of each change, and the reason for the change. In effect, the audit trail acts as a "chain-of-custody" record for the data base.

Software Design

No commercially available software was found that could be cost-effectively modified to meet the validation requirements of the CEL facility. (Other researchers have also found a lack of "off-the-shelf" software that facilitates validation of health related data acquired for research purposes (*5, 6*).) Consequently, a program called CLEANUP was written to act as a "data validation nucleus." The design of CLEANUP is described in this section. The use of CLEANUP in conjunction with commercially available data-base management programs and statistical packages to form a validation system of adequate power and flexibility is described in subsequent sections.

CLEANUP Inputs

1. Data Records. CLEANUP operates on fixed-length, ASCII ("card image") records stored in a user-specified sequential data file. (The use of ASCII records allows CLEANUP to be easily integrated with off-the-shelf statistical packages and data-base management programs.) The first three-character field of each record must be numeric and defines the "Record Type." Any number of data records of different Record Types may be contained in random order within a single data file.

2. Data Dictionary. The format for each Record Type is specified by a user generated "Data Dictionary" which identifies each field by an alphanumeric name, followed by the field length, field type (alphanumeric, decimalnumeric, or integral numeric) and expected range of the data.

3. Validation Criteria File. Each Record Type has an associated Validation Criteria File defining the minimum and maximum permissible values for each field.

4. Commands. CLEANUP executes commands entered from a CRT terminal keyboard or obtained from a Command File. Three groups of commands are defined:

Viewing Mode commands. Viewing Mode commands are entered from a keyboard and allow the user to display selected records on a CRT screen.

These displayed records can then be modified, deleted, or printed in a user-friendly format to permit more careful inspection. The data records selected for viewing are defined by Search Commands (see below).

Batch Mode commands. Unlike the Viewing Mode commands, which act only on the data displayed, Batch Mode commands are applied to all records in the data file that meet criteria defined by the Search commands (see below). Batch Mode commands may also be executed from a Command File, permitting the computer to perform routine data validation activities without operator intervention. Batch Mode commands are available for deleting records, modifying data within a record, printing records, flagging data within a record which are outside the upper and lower limits provided in the Validation Criteria File, and restoring flagged data to their preflagged values.

Search commands. Search commands, obtained from a keyboard or the Command File, allow the selection of records within a data file. The general form of a search command is

field name ⟨operator⟩ value

where field name is the name assigned to the field by the Data Dictionary ⟨operator⟩ is an arithmetic operator (=, >, <, ≥, ≤, and the negation of these symbols), and value is an alphanumeric constant (e.g., 100 or SMITH). (A typical search command would be FVC > 5, which would select all records with a spirometric forced vital capacity greater than 5 liters.) Up to 10 Search commands may be specified. The Viewing and Batch Mode commands are applied only to those records which satisfy the current set of search criteria.

CLEANUP Outputs

1. Data Record Modification. CLEANUP can be used to modify any field within a data record according to commands entered by the user or executed from the Command File.

2. Data Reports. CLEANUP can provide CRT displays or hard-copy reports of selected data in a user-friendly format.

3. Audit Trail. If a CLEANUP command makes any change to a record, a before-and-after image of the record is written to the Audit Trail associated with each data file. The date and time the change was made as well as the name of the person making the change are also written to the Audit Trail.

CLEANUP Design

CLEANUP is a modularly designed program written in AIS-PL/I (Applied Information Systems, Chapel Hill, N.C.) which runs on DEC (Digital Equipment Corp., Maynard, Mass.) PDP-11 family processors. Program design is summarized in Fig. 1 (numbers in brackets refer to corresponding numbers in Fig. 1).

Data Record formats are defined in a Data Dictionary [1] which is read and interpreted by GETDIC [2]. Search Criteria obtained from the Keyboard [3] or

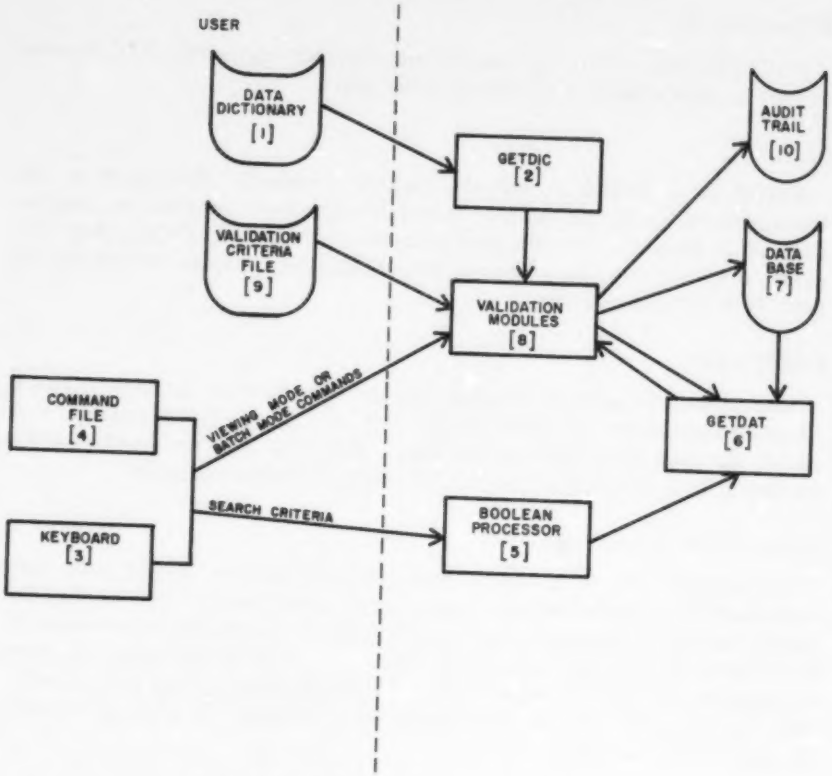

FIG. 1. Block diagram of program CLEANUP.

Command File [4] are interpreted by a Boolean Processor [5] and used by GETDAT [6] to acquire selected records from the Data Base [7] as commanded by the Validation Modules [8].

Validation Modules are executed as required by commands received from the Keyboard [3] or Command File [4] and perform the actual modifications to (flagging, deleting, restoring or adding data, etc.) or actions on (printing, sorting) the data records according to the commands they receive and the format for each Record Type as obtained from GETDIC. If necessary, the Validation Module also reads the Validation Criteria File [9] for the appropriate Record Type. Whenever a data record is modified, a before and after image of the record is written to the Audit Trail [10].

CEL VALIDATION SYSTEM

CLEANUP is used as the nucleus of the CEL validation system which includes the support software described below.

DATATRIEVE

DATATRIEVE, a data-base management system supplied by DEC, is used to provide data reports to the PI and other users.

BMDP

BMDP is a statistical analysis package (originally developed at the University of California at Los Angeles) especially designed to analyze biomedical data. The version used to assist in validation of CEL data was modified by Software Development Inc. (Middlebury College, Vt.) to run on DEC PDP-11 family machines.

EASYENTRY

Manual data records are entered into the data base using EASYENTRY (Applied Information Systems, Chapel Hill, N.C.), a software package that allows the user to design and use forms for data entry on standard CRT terminals. It is designed to run on DEC PDP-11 family computers.

Programs for Interrecord Check

CLEANUP will not compare data contained in one record with data contained in a second record. These "interrecord checks" are handled by special purpose software programs which generate Command Files containing CLEANUP Batch Mode commands. These Batch Mode commands are then executed by CLEANUP, which makes the actual change to the data records. (For data integrity reasons, no program but CLEANUP is allowed to make changes to data records.)

Validation of CEL Data

Data records are obtained from the PDAS, the Aerosol Control System, the Gas Control System, and from the use of EASYENTRY. A PDAS record includes the Record Type identifier, subject number (in order to comply with the Privacy Act names are never used), protocol identification number, date and time of data acquisition, condition codes defining the subjects pollutant exposure prior to the test, and the results of the physiological test. The Gas and Aerosol Systems each produce two types of data records. Gas and Aerosol Control Records include a Record Type identifier, protocol identification number, date and time of acquisition, the concentration of the pollutant being measured, and status information associated with the gas or aerosol analyzer making the measurement. In addition, calibration of a gas or aerosol analyzer results in the production of a Calibration Record which includes the analyzer calibration constants along with the date and time of calibration.

The purpose of the validation process is to identify the subset of data outside the PI selected criteria limits. These "suspicious" records are then delivered to

the PI in a report along with all related data (e.g., data from repetition of the same pulmonary function test) needed for him to decide whether the data are valid or invalid. If the PI disagrees with the designation of parameters by the computer as valid or unvalid, he/she may modify the data base by using CLEANUP in an interactive mode. Integrity of the data is maintained by utilizing a password system that permits only the PI or a designated representative to modify the data values.

Use of the validation system is described below for three CEL data types.

Physiological Data Validation

Figure 2 illustrates the procedures used to validate physiological data records (numbers in brackets refer to the figure). The original physiological data records [1] are examined by program DESCHK [2]. DESCHK performs interrecord checks to ensure that the physiological tests are executed in the proper order and repeated the required number of times as defined by the Protocol Design File [3]. Any deviation from the original protocol design (wrong date or time, wrong number of trials, missing trials) results in an Audit Trail entry [4]. Information in the Audit Trail is used, along with PDAS Operator Error Report Sheets [6], to correct the data records using CLEANUP [5] in the interactive mode. (PDAS Operator Error Report Sheets allow an operator to log any deviations from the original protocol during the performance of the experiment due to equipment malfunction or mistakes by the instrumentation operator or subject.) Typical data base corrections include manual entry of data (obtained from log books) lost due to computer malfunction and correction of dates and times entered incorrectly by PDAS operators.

The corrected data records [7] are processed by CLEANUP [8] to flag out-of-limits (not physically reasonable) physiological data as defined by the Validation Criteria File [9] (e.g., spirometric vital capacities less than 3 liters or greater than 8 liters are flagged). Before and after copies of each flagged record are written to the Audit Trail [14].

The data records [10] are then subjected to statistical tests using BMDP [11] to identify possible outliers [12]. The types of tests used depend on the experimental design and are selected by the PI. A typical test would identify all physiological parameters more than three standard deviations from their group means for each experimental condition (e.g., clean air and pollutant exposure). Suspicious parameters are flagged using CLEANUP [13].

DATATRIEVE [16] is used in conjunction with the Validated Data Base [15] to prepare data reports to be delivered to the PI [17]. (The format of the reports is dependent upon the experimental design and is determined by the PI prior to the beginning of the experiment.) At this point, the PI can (using CLEANUP interactively) utilize his scientific judgment to make further changes to the data base or restore data previously flagged by the automated validation procedures. All changes made by the PI are, of course, recorded in the Audit Trail. Under

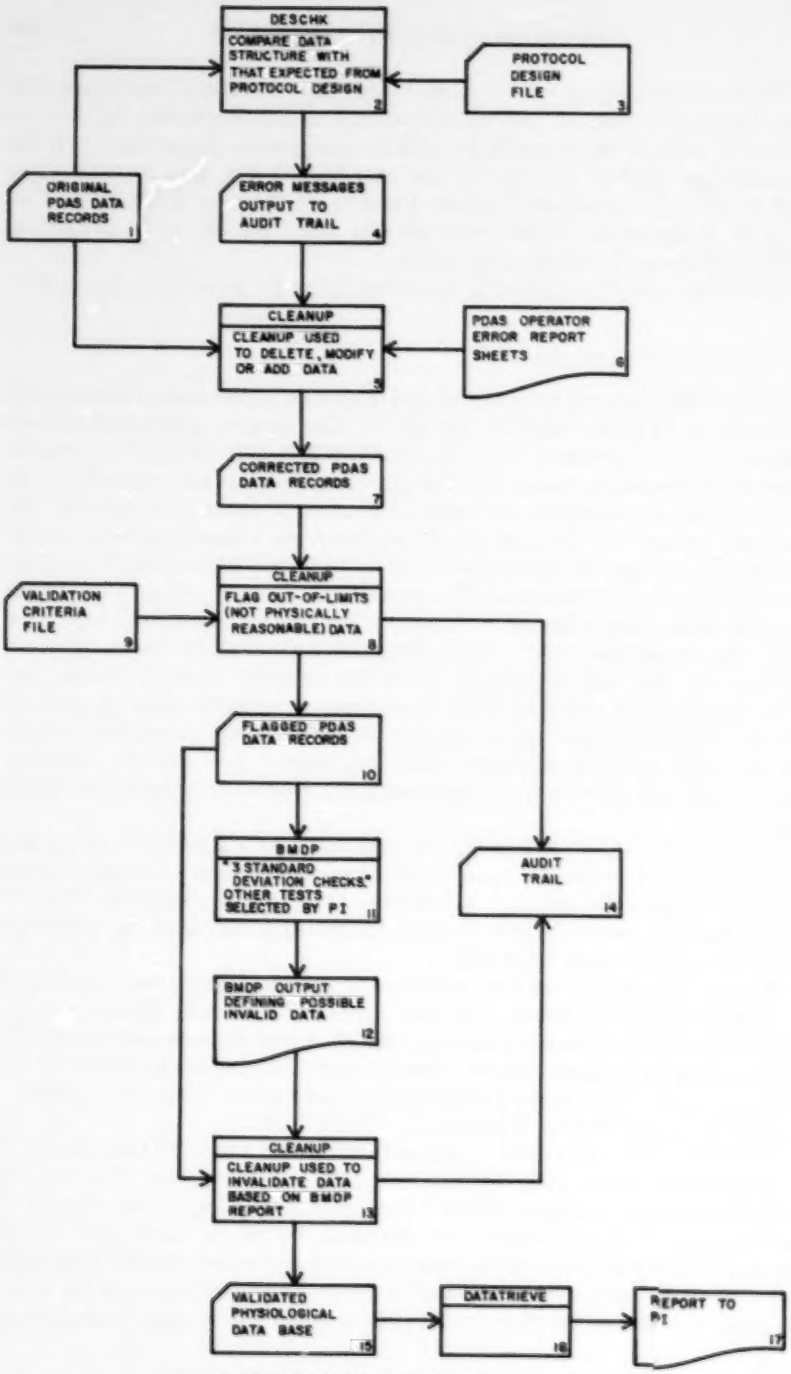

FIG. 2. Physiological data record validation scheme.

extreme circumstances, the PI may even change the screening criteria and request that the data be revalidated.

Gas Control System Data Validation

Figure 3 illustrates the procedures for validating data from the Gas Control System (numbers in brackets refer to the figure). The original records [1] are first processed by DESCHK [2], which checks that data records exist for the entire time subjects were exposed to pollutants according to the Protocol Design File [3]. Error messages identifying missing data are written to the Audit Trail [4] and used in conjunction with CLEANUP [5] to correct any errors in date and time and to manually enter (from logbooks) any control data lost due to computer malfunction.

The data records from the corrected data base [6] are then sorted [7] to separate the Control Records [8] and the Calibration Records [14].

The Control Records are examined for out-of-limits gas analyzer status information [10] (e.g., incorrect solenoid valve positions, improper flow of operating gases) using CLEANUP in conjunction with a Validation Criteria File [9]. Before and after images of each modified record are written to an Audit Trail [19]. The results [11] are once more processed by CLEANUP [12], which flags all gas concentrations outside the limits specified by the Validation Criteria File [9].

The Calibration Records [14] are first processed by CLEANUP [15] to identify any calibration constants that are outside acceptable limits as defined by a Validation Criteria File [16]. The validated records [17] are then processed by CALDRFT [18], which detects calibration drifts outside of acceptable limits and produces Batch Mode commands [19] used by CLEANUP [20] to flag Control Records [13] acquired by instruments with unacceptable calibration drift. The data are then delivered to the PI [21] for final inspection.

Manually Entered Data

Manual data (mainly questionnaire replies by the subjects about their medical symptoms) are entered into the data base using EASYENTRY. Real-time validation consists of ensuring that responses are relevant to the question. Inappropriate answers are automatically rejected by EASYENTRY and the subject is reprompted.

All subject symptom questionnaire data are entered directly into the data base by the subjects themselves from terminals located within the exposure chambers. This increases data integrity and validity since the errors in transcribing data from one format to another, or keying data from paper forms into the computer, are significant. A preliminary study made by one of the authors (RWS) indicated that almost 10% of data transcribed from one format to another were copied with errors.

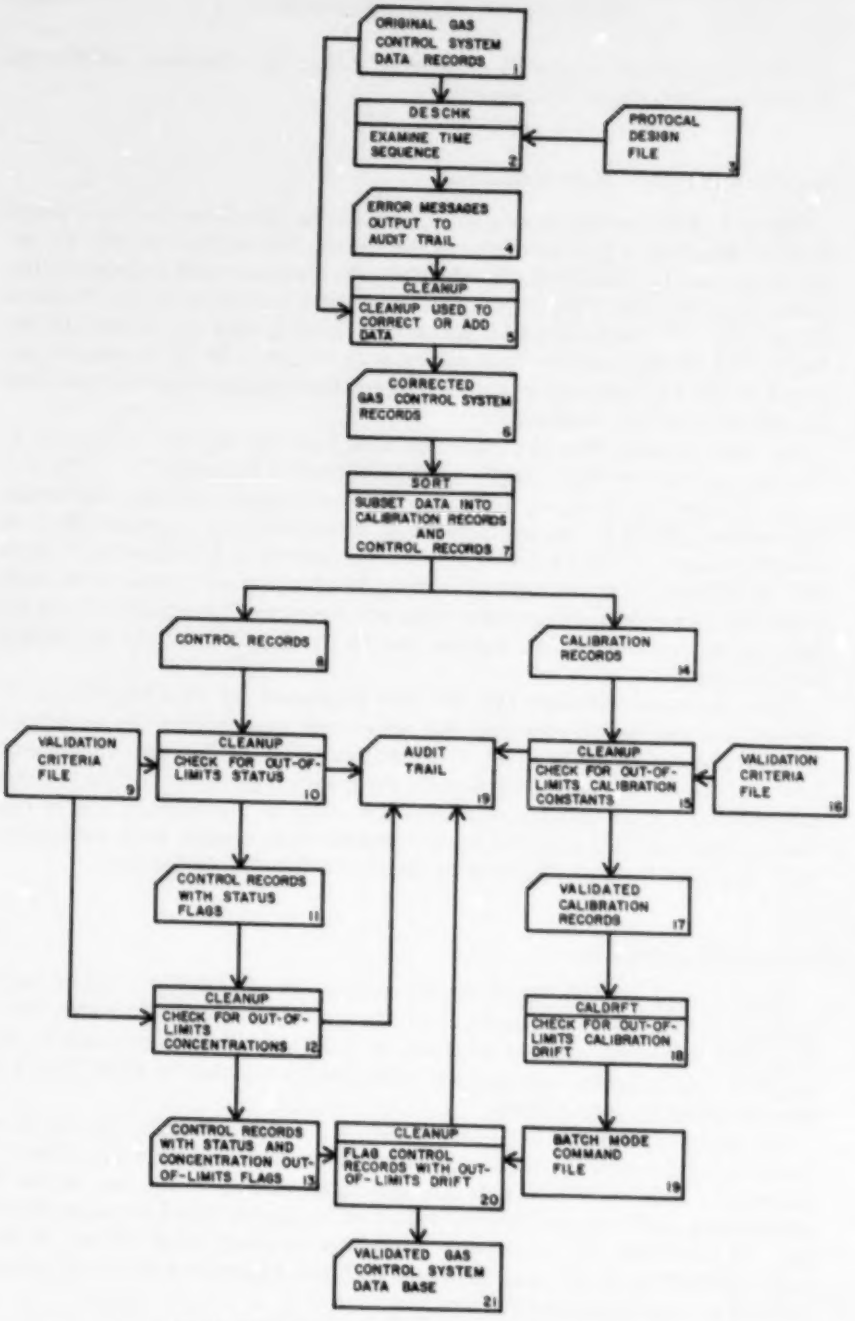
ORIGINAL GAS CONTROL SYSTEM DATA RECORDS 1
DESCHK
EXAMINE TIME SEQUENCE 2
PROTOCAL DESIGN FILE 3
ERROR MESSAGES OUTPUT TO AUDIT TRAIL 4
CLEANUP
CLEANUP USED TO CORRECT OR ADD DATA 5
CORRECTED GAS CONTROL SYSTEM RECORDS 6
SORT
SUBSET DATA INTO CALIBRATION RECORDS AND CONTROL RECORDS 7
CONTROL RECORDS 8
CALIBRATION RECORDS 14
VALIDATION CRITERIA FILE 9
CLEANUP
CHECK FOR OUT-OF-LIMITS STATUS 10
AUDIT TRAIL 19
CLEANUP
CHECK FOR OUT-OF-LIMITS CALIBRATION CONSTANTS 15
VALIDATION CRITERIA FILE 16
CONTROL RECORDS WITH STATUS FLAGS 11
VALIDATED CALIBRATION RECORDS 17
CLEANUP
CHECK FOR OUT-OF-LIMITS CONCENTRATIONS 12
CALDRFT
CHECK FOR OUT-OF-LIMITS CALIBRATION DRIFT 18
CONTROL RECORDS WITH STATUS AND CONCENTRATION OUT-OF-LIMITS FLAGS 13
CLEANUP
FLAG CONTROL RECORDS WITH OUT-OF-LIMITS DRIFT 20
BATCH MODE COMMAND FILE 19
VALIDATED GAS CONTROL SYSTEM DATA BASE 21

FIG. 3. Gas control system data record validation scheme.

Conclusion

The validation system described reduces the volume of data which must be carefully examined by the PI, making it possible to effectively validate data from large studies. However, the computerized system is only an aid to the PI who must have ultimate responsibility for the integrity of the data base. In particular, the computerized validation system does not pretend to substitute for the pattern recognition ability of the trained scientist.

It can be argued that a validation system of this complexity is unnecessary in a clinical research environment, because the percentage of invalid data collected by knowledgeable investigators is likely to be small and will not affect the conclusions of the study. Although this is undoubtedly true, the detection of a small percentage of invalid or improperly assigned data during a legal challenge could result in a lack of confidence in the entire data base and could require that the data from the experiment be completely reanalyzed.

An important limitation of the CEL validation system is its inability to perform interrecord checks. Modification of CLEANUP to perform limited interrecord checking is under investigation. Interrecord checks requiring data from more than two records will probably continue to be handled by special purpose programs.

The validation system developed for the CEL facility has proven to be flexible and has been used to validate data from areas unrelated to health research. Data associated with an atmospheric visibility study and a study of the size distribution of oil droplets emitted from an offshore oil platform were successively validated and analyzed using this system. CLEANUP has also been used to develop a rudimentary but useful inventory control system. The source code for CLEANUP is available from the authors upon request.

Acknowledgment

This work was performed under EPA Contract 68-02-3492.

References

1. Strong, A. A. "Description of the CLEANS Human Exposure System." U.S. EPA, EPA-600/1-78-064, 1978.
2. Crider, W. L., Landis, D. A., and Weaver, F. H. Computer-controlled human exposure facilities utilizing water soluble aerosols. *In* "Generation of Aerosols and Facilities for Exposure Experiment" (K. Willeke, Ed.), pp. 493–515. Ann Arbor Science Pub., Ann Arbor, Mich. 1980.
3. Glover, D. E., Berntsen, J. H., Crider, W. L., and Strong, A. A. Design and performance of a system to control concentrations of common gaseous air pollutants within environmental laboratories used for human exposure studies. *J. Environ. Sci. Health, P. A* **16** (1981).
4. "Quality Assurance Handbook for Air Pollution Measurement System," Vol. I, Principles. U.S. EPA, EPA-600/9-76-005.
5. Wolfenson, L. B., and Worth, N. An approach to health data validation. *Comput. Biomed. Res.* **13**, 501 (1980).
6. MacLaughlin, D. S. A validation program nucleus. *Comput. Prog. Biomed.* **11**, 43 (1980).